Memory Bandits: a Bayesian approach for the Switching Bandit Problem

Réda Alami, Odalric Maillard, Raphael Féraud

Memory Bandits: a Bayesian approach for the Switching Bandit Problem

Réda Alami
Orange Labs
reda1.alami@orange.com

Odalric Maillard
INRIA (SequeL Team)
odalric.maillard@inria.fr

Raphael Féraud
Orange Labs
raphael.feraud@orange.com

Abstract

The Thompson Sampling exhibits excellent results in practice and it has been shown to be asymptotically optimal. The extension of Thompson Sampling algorithm to the Switching Multi-Armed Bandit problem, proposed in [13], is a Thompson Sampling equiped with a Bayesian online change point detector [1]. In this paper, we propose another extension of this approach based on a Bayesian aggregation framework. Experiments provide some evidences that in practice, the proposed algorithm compares favorably with the previous version of Thompson Sampling for the Switching Multi-Armed Bandit Problem, while it outperforms clearly other algorithms of the state-of-the-art.

1 Introduction

We consider the non-stationary multi-armed bandit problem with a set $\mathcal{K}$ of K independent arms. At each round t, the agent chooses an action $k_t \in \mathcal{K}$ and observes a reward x_{k_t}. In a stationary environment, the agent has to explore to find the best arm and to exploit it to maximize his gain. Efficient algorithms [4, 7, 11] have been proposed for handling the so-called exploration-exploitation dilemma. In an evolving environment, the best arm can change during time and hence the agent has to explore more. The adversarial bandits handles non-stationary environments by considering that a sequence of deterministic rewards is chosen in advance by an oblivious adversary. Rate optimal algorithms have been proposed to find the best arm or the best sequence of arms of the run in [5, 14, 3]. Another approach for handling non-stationary environment is to consider that the rewards are generated by an unknown stochastic process that evolves during time. In the switching bandit problem [10, 9, 2], the mean rewards of arms change abruptly. In comparison to the adversarial approach, the advantage of non-stationary stochastic approach is that with some mild assumptions, stochastic algorithms, which are more efficient in practice than adversarial algorithms, can be used. For practical and theoretical reasons, the recent years have seen an increasingly interest for the oldest bandit algorithm, the Thompson Sampling [15]. It exhibits excellent results in practice [8], while it is asymptotically optimal [11]. In [13], the authors propose an adaptation of the Thompson Sampling to the switching bandit problem. To be consistent with the Bayesian approach, rather than using a frequentist drift detector used in [10, 2], the authors use a Bayesian online change point detection [1] combined with the Thompson Sampling algorithm. In this paper, we propose a similar approach of [13] which is based on a Bayesian aggregation of a growing number of experts seen as learners. Our approach compares favorably with the one of [13].

2 Problem formulation

Let us consider an agent facing a non-stationary multi-armed bandit problem with a set $\mathcal{K} = \{1, ..., K\}$ of K independent arms. At each round $t \in [\![1, T]\!]$, the agent chooses to observe one of the K possible actions. When playing the arm k_t at time t, a reward x_{k_t} is received, where $x_{k_t} \sim \mathcal{B}(\mu_{k_t,t})$ is a random variable drawn from a Bernoulli distribution of expectation $\mu_{k,t}$. Let $\mu_t^\star = \max_{k\in\mathcal{K}} \{\mu_{k,t}\}$ denotes the best expected reward at round t, $k_t^\star = \arg\max_{k\in\mathcal{K}} \{\mu_{k,t}\}$ the best arm at round t, k_t the action chosen by the decision-maker at time t and x_{k_t} the reward obtained at the same time.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Changes in the Bernoulli distributions expectations It should be noted that $\mu_{k,t}$, i.e. the reward mean of arm k at time t changes over time according to a global abrupt switching model parametrized with an unknown hazard function $h(t) \in [0,1]$ assumed to be a constant $h(t) = \rho$ such that:

$$\mu_{k,t} = \begin{cases} \mu_{k,t-1} & \text{with probability } 1-\rho \\ \mu_{new} \sim \mathcal{U}(0,1) & \text{with probability } \rho \end{cases} \tag{1}$$

When the behavior's environment is modeled by equation (1) for all $k \in \mathcal{K}$, the problem setting is called a *Global Switching Multi-Armed Bandit* (GS-MAB), i.e. when a switch happens *all* arms change their expected rewards. Where changes occur independently for each arm k (i.e. arms change points are independent from an arm to another), the problem setting is called a *Per-arm Switching Multi-Armed Bandit*. For the sake of clarity, in the following we will focus on GS-MAB.

Sequence of change points It should be noted that for each GS-MAB, it exists a non-decreasing change points sequence of length Υ_T denoted by $(\tau_\kappa)_{\kappa \in [\![1, \Upsilon_T+1]\!]} \in \mathbb{N}^{\Upsilon_T+1}$ where:

$$\begin{cases} \forall\ \kappa \in [\![1, \Upsilon_T]\!],\ \forall\ t \in \mathcal{T}_\kappa = [\![\tau_\kappa + 1, \tau_{\kappa+1}]\!],\ \forall\ k \in \mathcal{K},\ \ \mu_{k,t} = \mu_{k,[\kappa]} \\ \tau_1 = 1 < \tau_2 < ... < \tau_{\Upsilon_T+1} = T \end{cases}$$

In this case, $\mu^\star_{[\kappa]} = \max_k \left\{\mu_{k,[\kappa]}\right\}$ denotes the highest expected reward at epoch $\mathcal{T}_\kappa$.

Pseudo Cumulative Regret for the switching environment In a switching environment, the pseudo cumulative regret $\mathcal{R}(T)$ up to time T is defined as the expected difference between the rewards obtained by our policy and those received by the *oracle* which always plays the best arm $k^\star_{[\kappa]}$ at each epoch $\mathcal{T}_\kappa$ such as:

$$\mathcal{R}(T) = \sum_{t=1}^{T} \mu^\star_t - \mathbb{E}\left[\sum_{t=1}^{T} x_{k_t}\right] = \sum_{\kappa=1}^{\Upsilon_T} \left(|\mathcal{T}_\kappa|\, \mu^\star_{[\kappa]} - \mathbb{E}\left[\sum_{t=\tau_\kappa+1}^{\tau_{\kappa+1}} x_{k_t}\right]\right)$$

3 Global Switching Thompson Sampling with Bayesian Aggregation

3.1 The Thompson Sampling algorithm

Unlike optimistic algorithms belonging to the UCB family [4], which are often based on confidence intervals, the Thompson Sampling deals with Bayesian tools by assuming a Beta prior distribution $\pi_{k,t=1} = \text{Beta}(\alpha_0, \beta_0)$ on each arm for some $\alpha_0, \beta_0 > 0$. Based on the rewards observed x_{k_t}, the posterior distribution $\pi_{k,t}$ is updated such as: $\pi_{k,t} = \text{Beta}\left(\alpha_{k,t} = \#(\text{reward} = 1) + \alpha_0, \beta_{k,t} = \#(\text{reward} = 0) + \beta_0\right)$. At each time, the agent takes a sample $\theta_{k,t}$ from each $\pi_{k,t}$ and then plays the arm $k_t = \arg\max_k \theta_{k,t}$. Formally, by denoting $D_{t-1} = \bigcup_{i=1}^{t-1} x_i$ the history of past rewards we write: $\theta_t = (\theta_{1,t}, ..., \theta_{K,t}) \sim \mathbb{P}(\theta_t | D_{t-1}) = \prod_{k=1}^{K} \pi_{k,t}$. Recently, the Thompson Sampling has been shown to be asymptotically optimal [11], i.e. the expectation of the pseudo-cumulative regret reaches the Lai and Robbins lower bound on regret in the stochastic Bernoulli multi-armed bandit setting [12].

3.2 Decision making based on Bayesian aggregation

Best achievable performance: the Thompson Sampling oracle Let $TS^\star$ denotes the oracle that knows exactly the change points τ_κ. It simply restarts a Thompson Sampling at these change points. Assume that Υ_T is the overall number of change points observed until T, then $TS^\star$ runs successively Υ_T Thompson Sampling processes starting at $\tau_\kappa + 1$ and ending at $\tau_{\kappa+1}$.

Notion of expert Let $t \in \mathbb{N}^\star$ and $i \in [\![1, t]\!]$. An expert $f_{i,t}$ is a Thompson Sampling procedure which has started at time i. The expert $f_{i,t}$ observes exactly $t - i$ rewards from the environment.

Expert Aggregation Like [13], to characterize the occurrence of changes, we use the expert $f_{i,t}$ as an index to access in the memory the parameters of the model created at time i. The computation of $\mathbb{P}(\theta_t | D_{t-1})$ is done by taking into account the distribution $w_{i,t} = \mathbb{P}\left(f_{i,t} | D_{t-1}\right)$ of the expert $f_{i,t}$ such as:

$$\mathbb{P}(\theta_t | D_{t-1}) = \sum_{i=1}^{t} \mathbb{P}(\theta_t | D_{t-1}, f_{i,t}) \mathbb{P}(f_{i,t} | D_{t-1}) \tag{2}$$

Then, unlike the sampling procedure used in [13], we build the index $\theta_{k,t}$ of arm k at time t by launching a Bayesian aggregation of a growing number of experts. The estimation of the expert distribution is done recursively according to the work of [1] where:

$$\underbrace{\mathbb{P}\left(f_{i,t} | D_{t-1}\right)}_{\texttt{Expert distribution at } t} \propto \sum_{i=1}^{t-1} \overbrace{\mathbb{P}(f_{i,t} | f_{i,t-1})}^{\texttt{change point prior}} \underbrace{\mathbb{P}(x_{k_t} | f_{i,t-1}, D_{t-2})}_{\texttt{Instantaneous gain}} \overbrace{\mathbb{P}(f_{i,t-1} | D_{t-2})}^{\texttt{Expert distribution at } t-1} \tag{3}$$

The change point prior $\mathbb{P}(f_{i,t}|f_{i,t-1})$ is naturally computed following equation (1):

$$\mathbb{P}(f_{i,t}|f_{i,t-1}) = (1-\rho)\,\mathbb{1}\,(i<t) + \rho\mathbb{1}\,(t=1) \tag{4}$$

Thus, the inference model takes the following form (Up to a normalization factor):

$$\begin{cases} \textit{Growth probability:} & \mathbb{P}(f_{i,t}|D_{t-1}) \propto (1-\rho)\cdot\mathbb{P}(x_{k_t}|f_{i,t-1},D_{t-2})\cdot\mathbb{P}(f_{i,t-1}|D_{t-2}) \\ \textit{change point probability:} & \mathbb{P}(f_{t,t}|D_{t-1}) \propto \rho\sum_{i=1}^{t-1}\mathbb{P}(x_{k_t}|f_{i,t-1},D_{t-2})\cdot\mathbb{P}(f_{i,t-1}|D_{t-2}) \end{cases} \tag{5}$$

It should be noted that $\mathbb{P}(x_{k_t}|f_{i,t-1},D_{t-2})$ is a Bernoulli distribution of expectation $\frac{\alpha_{k_t,i,t-1}}{\alpha_{k_t,i,t-1}+\beta_{k_t,i,t-1}}$, where $\alpha_{k_t,i,t-1}$ and $\beta_{k_t,i,t-1}$ are the hyper-parameters of the arm k_t learned by the expert $f_{i,t-1}$. Let $l_{i,t-1}$ denotes the instantaneous logarithmic loss associated to the forecaster $f_{i,t-1}$ such as: $\mathbb{P}(x_{k_t}|f_{i,t-1},D_{t-2}) = \exp(-l_{i,t-1})$, then Algorithm 1 provides us an index prediction of each arm k at time t based on a Bayesian aggregation of the available experts.

Algorithm 1 Bayesian Aggregation with a growing number of experts

for $t = 2, ...$ **do**

-1- Update: $\forall i \in \{1,...,t-1\}\ \ w_{i,t} \leftarrow (1-\rho)\,w_{i,t-1}\exp(-l_{i,t-1})$

-2- Create new expert starting at t: $w_{t,t} \leftarrow \rho\sum_{i=1}^{t-1} w_{i,t-1}\exp(-l_{i,t-1})$

-3- Predict arm index: $\forall\ k \in \mathcal{K}\ \ \theta_{k,t} \leftarrow \frac{\sum_{i=1}^{t-1}\theta_{k,i,t}w_{i,t}}{\sum_{i=1}^{t-1}w_{i,t}}$ where: $\theta_{k,i,t} \sim \mathrm{Beta}(\alpha_{k,i,t},\beta_{k,i,t})$

Finally, by plugging the Bayesian aggregation into the formalism of [13], we get the Global Switching Thompson Sampling with Bayesian Aggregation (Global-STS-BA), described in Algorithm 2.

Algorithm 2 Global Switching **T**hompson **S**ampling with **B**ayesian **A**ggregation

procedure GLOBAL-STS-BA($\mathcal{K}, T, \alpha_0, \beta_0, \rho$)
  $t \leftarrow 1, w_{1,t} \leftarrow 1$, and $\forall k \in \mathcal{K}\ \alpha_{k,1,t} \leftarrow \alpha_0,\ \beta_{k,1,t} \leftarrow \beta_0$ ▷ Initializations
  for $t \leqslant T$ **do** ▷ Interaction with environment
    $k_t \leftarrow$ CHOOSEARM $(\{w\}^t, \{\alpha\}^t, \{\beta\}^t)$
    $x_{k_t} \leftarrow$ PLAYARM(k_t) ▷ Bernoulli trial
    $\{w\}_{t+1} \leftarrow$ UPDATEEXPERTWEIGHT $(\{w\}_t, \{\alpha\}_{k_t,t}, \{\beta\}_{k_t,t}, x_{k_t}, \rho)$
    $\{\alpha\}_{t+1}, \{\beta\}_{t+1} \leftarrow$ UPDATEARMMODEL$(\{\alpha\}_t, \{\beta\}_t, x_{k_t}, k_t)$

procedure CHOOSEARM$(\{w\}_t, \{\alpha\}_t, \{\beta\}_t)$
  $\forall\ k \in \mathcal{K}\ \forall\ i \in [\![1,t]\!]\ \ \theta_{k,i,t} \leftarrow \mathrm{Beta}(\alpha_{k,i,t},\beta_{k,i,t})$
  return $\arg\max_k \sum_{i\in[\![1,t]\!]} \frac{w_{i,t}}{\sum_{j\in[\![1,t]\!]} w_{j,t}}\,\theta_{k,i,t}$ ▷ Bayesian aggregation

procedure UPDATEEXPERTWEIGHT$(\{w\}_t, \{\alpha\}_{k_t,t}, \{\beta\}_{k_t,t}, x_{k_t}, \rho)$
  $l_{i,t} \leftarrow -x_{k_t}\log\left(\frac{\alpha_{k_t,i,t}}{\alpha_{k_t,i,t}+\beta_{k_t,i,t}}\right) - (1-x_{k_t})\log\left(\frac{\beta_{k_t,i,t}}{\alpha_{k_t,i,t}+\beta_{k_t,i,t}}\right)\ \forall\ i \in [\![1,t]\!]$
  $w_{i,t+1} \leftarrow (1-\rho)\,w_{i,t}\exp(-l_{i,t})\ \forall\ i \in [\![1,t]\!]$ ▷ Increasing the size of expert $f_{i,t}$
  $w_{t+1,t+1} \leftarrow \rho\sum_i w_{i,t}\exp(-l_{i,t})$ ▷ Creating new expert starting at $t+1$
  return $\{w\}_{t+1}$

procedure UPDATEARMMODEL$(\{\alpha\}_t, \{\beta\}_t, x_{k_t}, k_t)$
  $\alpha_{k_t,i,t+1} \leftarrow \alpha_{k_t,i,t} + \mathbb{1}\,(x_{k_t}=1)\ \forall\ i \in [\![1,t]\!]$
  $\beta_{k_t,i,t+1} \leftarrow \beta_{k_t,i,t} + \mathbb{1}\,(x_{k_t}=0)\ \forall\ i \in [\![1,t]\!]$
  $\alpha_{k,t+1,t+1} \leftarrow \alpha_0,\ \ \beta_{k,t+1,t+1} \leftarrow \beta_0\ \ \forall\ k \in \mathcal{K}$ ▷ Initializing new expert
  return $\{\alpha\}_{t+1}, \{\beta\}_{t+1}$

Global-STS-BA In the global switching setting, when a switch occurs, all arms change their expected pay-off at the same time. It should be noted that the data from all plays collaborate in the posterior of the expert distribution estimation (UPDATEEXPERTWEIGHT). The experts tell us how much previous observed data can be used in the arm indexes prediction, i.e. when a switch occurs all past data have not to be taken into account in the arm characterizations because of their obsolete information. The change point detection concept is based on a tracking of the optimal expert (see Appendix B.1). Indeed, the total mass of the expert distribution $w_{i,t}$ tends to focus around the *optimal* expert $f_{\tau_\kappa,t}$ i.e. the expert starting at the most recent change point τ_κ and corresponds to the most appropriate characterization of the environment. The Bayesian aggregation exploits this

concentration to highlight the contribution of the most appropriate experts (starting around the most recent change point τ_κ) in the arm indexes prediction. The concentration around $f_{\tau_\kappa,t}$ is possible thanks to the inference model of equation (5). In fact, when a change occurs the instantaneous gain $\mathbb{P}\left(x_{k_t}|f_{i,t-1}, D_{t-2}\right)$ of all experts starting before the change point suddenly fall down because of their wrong estimation of the environment, giving the advantage to the experts newly created while annihilating the former ones (see Appendix B.1). At this point, we deviate from [13] and instead of sampling the expert distribution and then sampling the arms, we index each arm k by launching an overall Bayesian aggregation of samples taken from the posterior distribution of arm k related to the hyper-parameters $\{\alpha\}_{k,t}, \{\beta\}_{k,t}$ (CHOOSEARM). Finally, the agent chooses to pull the arm with highest index. This process allows us to avoid the sampling noise induced by Global-STS and by this way the model chosen at time t tends to better fit the unknown environment (figure 1).

4 Experiments

In all the experiments, we consider a GS-MAB of three arms observing three change points occurring at each 1000 rounds. Experiments are run 100 times. The parameters of the state-of-the art algorithms are chosen to be experimentally optimal. Exp3, Exp3P and Exp3S [5] are launched with an exploration rate ($\gamma = 5\%$). We run Exp3R [2], Rexp3 [6] and SW-UCB [9] respectively with $H = 1000$, $\Delta_T = 1000$ and $\tau = 500$. Global-STS [13] and Global-STS-BA are outperforming the well parametrized state of art non stationary MAB algorithms (figure 1).
Replacing the expert distribution sampling used in [13] with the Bayesian Aggregation allows us to obtain performances challenging those of the Thompson Sampling Oracle (figure 1).

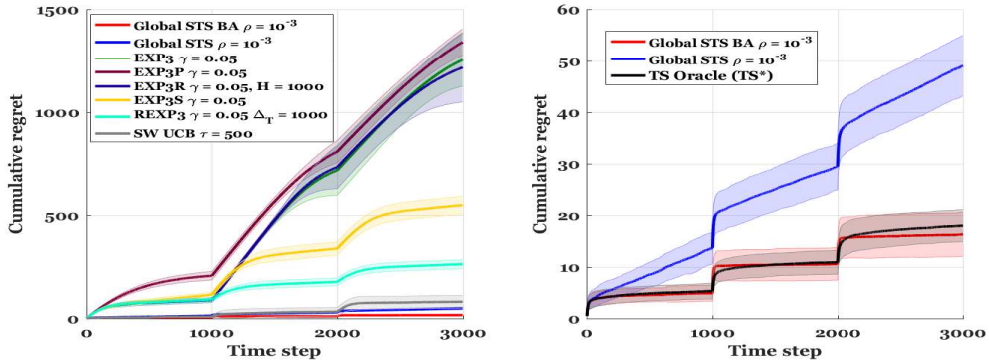


Figure 1: Overall comparison with the non-stationary state of art and the TS Oracle.

5 Conclusion and future works

We have proposed Global-STS-BA: an extension of the Thompson Sampling for the Switching Bandit Problem based on a Bayesian aggregation framework. From the experiments, the proposed algorithm compares favorably with the previous version of the Global Switching Thompson Sampling [13], outperforming clearly other algorithms of the state-of-the-art. It is worth noting that Global-STS-BA challenges the Thompson sampling oracle, an oracle which already knows the change points. These results arise from the fact that Global-STS-BA is based on the Bayesian concept of tracking the best experts which allows us to catch efficiently the change points. The proposed algorithm is obviously extended to the *Per-arm Switching Multi-Armed Bandit* by allowing an expert distribution per arm (see Appendix A). The next step of this work is to analyze the Global-STS-BA in term of regret.

Acknowledgments

This work has been supported by the French Agence Nationale de la Recherche (ANR), under grant ANR-16- CE40-0002 (project BADASS).

References

[1] Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

[2] Robin Allesiardo and Raphaël Féraud. Exp3 with drift detection for the switching bandit problem. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–7. IEEE, 2015.

[3] Robin Allesiardo, Raphaël Féraud, and Odalric-Ambrym Maillard. The non-stationary stochastic multi-armed bandit problem. *International Journal of Data Science and Analytics*, pages 1–17, 2017.

[4] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[5] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, January 2003.

[6] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 199–207, Cambridge, MA, USA, 2014. MIT Press.

[7] Olivier Cappé, Aurélien Garivier, Rémi Maillard, Odalric-Ambrym Munos, and Gilles Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

[8] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *NIPS*, pages 2249–2257, 2011.

[9] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

[10] Cédric Hartland, Sylvain Gelly, Nicolas Baskiotis, Olivier Teytaud, and Michèle Sébag. Multi-armed bandit, dynamic environments and meta-bandits. 2006.

[11] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer, 2012.

[12] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

[13] Joseph Mellor and Jonathan Shapiro. Thompson sampling in switching environments with bayesian online change point detection. *CoRR, abs/1302.3721*, 2013.

[14] Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 3168–3176, Cambridge, MA, USA, 2015. MIT Press.

[15] William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

[16] Ryan Turner, Yunus Saatci, and Carl Edward Rasmussen. Adaptive sequential Bayesian change point detection. In *Advances in Neural Information Processing Systems (NIPS): Temporal Segmentation Workshop*, 2009.

A Extension to the per-arm switching setting

A.1 Per-Arm Switching Thompson Sampling with Bayesian Aggregation (Per-Arm-STS-BA)

In the per-arm switching setting, since the changes occur independently from an arm to another, each arm k will have its own expert distribution denoted by $\{w\}_k^t$. To estimate $\{w\}_k^t$ at each time step t, we need to build some recursive message-passing algorithm. The main difference between the global and the per-arm version of the message-passing is that the instantaneous gain will be used in the update of the expert distribution associated to the pulled arm k_t. For the other arms not pulled, the instantaneous gain won't intervene in the expert distribution [13]. More formally, if we denote by $f_{k,i,t}$ the expert associated to the arm k at time t which has been introduced at time i, then the expert distributions are updated as follow:

$$\mathbb{P}(f_{k,i,t}|D_{t-1}) \propto \begin{cases} \sum_{i=1}^{t-1} \mathbb{P}\left(f_{k,i,t}|f_{k,i,t-1}\right) \mathbb{P}(x_{k_t}|f_{k,i,t-1}, D_{t-2}) \mathbb{P}(f_{k,i,t-1}|D_{t-2}) & \text{if } k \text{ is pulled} \\ \sum_{i=1}^{t-1} \mathbb{P}\left(f_{k,i,t}|f_{k,i,t-1}\right) \mathbb{P}(f_{k,i,t-1}|D_{t-2}) & \text{otherwise} \end{cases}$$

Then, we model the changes of the arm k with a constant switching rate γ_k such that:

$$\mathbb{P}\left(f_{k,i,t}|f_{k,i,t-1}\right) = \begin{cases} 1-\rho_k & \text{if } i < t \\ \rho_k & \text{if } i = t \end{cases}$$

Then, following the inference model used in the global switch, we deduce the following expert distributions update rules (Up to a normalization factor):

$$\textit{Growth probability:} \qquad \mathbb{P}\left(f_{k,i,t}\right) \propto \begin{cases} (1-\rho_k)\mathbb{P}(x_{k_t}|f_{k,i,t-1})\mathbb{P}(f_{k,i,t-1}) & \text{if } k \text{ is pulled} \\ (1-\rho_k)\mathbb{P}(f_{k,i,t-1}) & \text{otherwise} \end{cases}$$

$$\textit{change point probability:} \qquad \mathbb{P}(f_{k,t,t}) \propto \begin{cases} \rho_k \sum_{i=1}^{t-1} \mathbb{P}(x_{k_t}|f_{k,i,t-1})\mathbb{P}(f_{k,i,t-1}) & \text{if } k \text{ is pulled} \\ \rho_k \sum_{i=1}^{t-1} \mathbb{P}(f_{k,i,t-1}) & \text{otherwise} \end{cases}$$

Notice that: $\mathbb{P}(x_{k_t}|f_{k,i,t-1})$ still a Bernoulli distribution of expectation $\frac{\alpha_{k_t,i,t-1}}{\alpha_{k_t,i,t-1}+\beta_{k_t,i,t-1}}$, where $\alpha_{k_t,i,t-1}, \beta_{k_t,i,t-1}$ are the hyper-parameters of arm k_t at time $t-1$ learned by the expert $f_{k_t,i,t-1}$ which is associated to the change point model of arm k_t. We define:

$$\mathbb{P}(x_{k_t}|f_{k_t,i,t-1}) = \exp\left(-l_{i,t-1}\right)$$

where: $l_{i,t-1}$ denotes the instantaneous logarithmic loss of the pulled arm k_t associated to the forecaster $f_{k_t,i,t-1}$. For convenience, we write: $w_{k,i,t} = \mathbb{P}(f_{k,i,t}|D_{t-1})$. Then, we naturally extend the Bayesian aggregation used in the global switching setting to the per-arm Bayesian Aggregation (Algorithm 3).

Algorithm 3 Per-arm Bayesian Aggregation

for $t = 2, ...$ **do**

-1- Update: $\forall i \in \{1, ..., t-1\}\ \forall\, k \in \mathcal{K}\ w_{k,i,t} \leftarrow (1-\rho_k)\, w_{k,i,t-1} \exp\left(-l_{i,t-1}\right)^{\mathbb{1}(k=k_t)}$

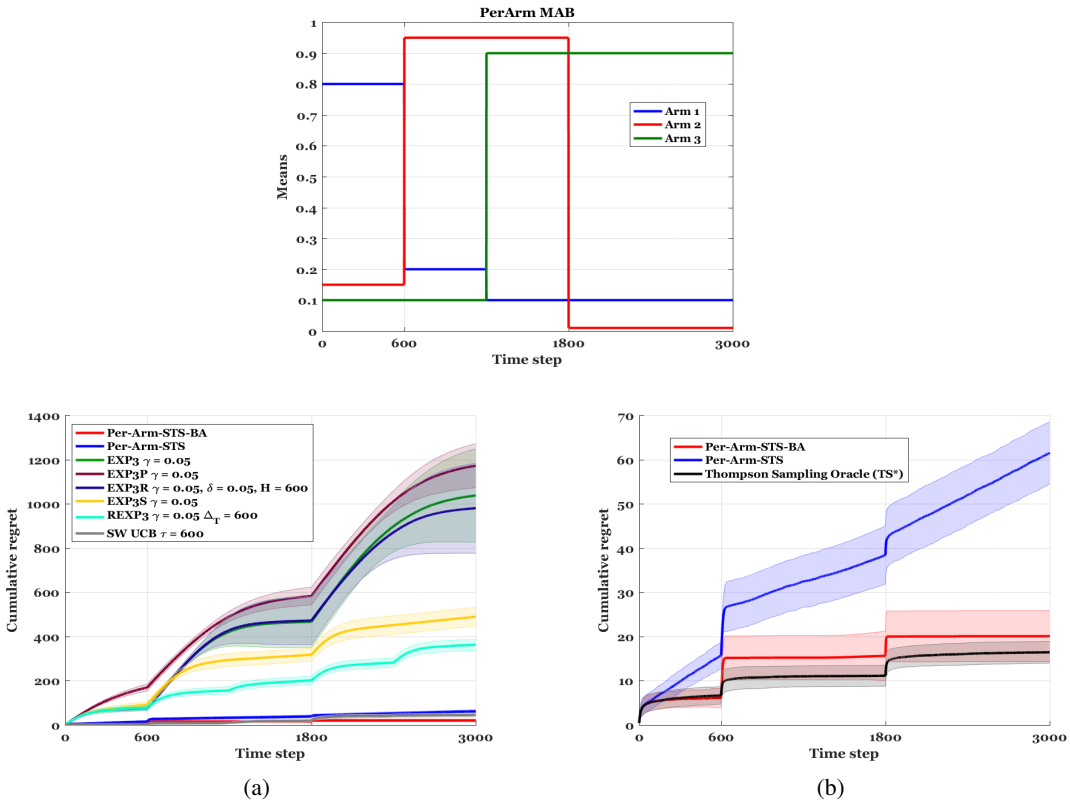
-2- Create new expert starting at t: $\forall\, k \in \mathcal{K}\ w_{k,t,t} \leftarrow \rho_k \sum_{i=1}^{t-1} w_{k,i,t-1} \exp\left(-l_{i,t-1}\right)^{\mathbb{1}(k=k_t)}$

-3- Predict arm index: $\forall\, k \in \mathcal{K}\ \theta_{k,t} \leftarrow \frac{\sum_{i=0}^{t-1} \theta_{k,i,t} w_{k,i,t}}{\sum_{i=0}^{t-1} w_{k,i,t}}$ where: $\theta_{k,i,t} \sim \text{Beta}\left(\alpha_{k,i,t}, \beta_{k,i,t}\right)$

Then, by plugging the per-arm Bayesian Aggregation into the formalism of the Global-STS-BA we get the *Per-arm Switching Thompson Sampling with Bayesian Aggregation* (Per-arm-STS-BA) described in Algorithm 4.

A.2 Experiments

In all the experiments, we consider a Per-Arm Switch MAB of three arms observing several change points. Experiments are run 100 times. Per-Arm-STS [13] and Per-Arm-STS-BA are outperforming the well parametrized state of art non stationary MAB algorithms (figure A.2). Exp3, Exp3P and Exp3S [5] are launched with an exploration rate ($\gamma = 5\%$). We run Exp3R [2], Rexp3 [6] and SW-UCB [9] respectively with $H = 600$, $\Delta_T = 600$ and $\tau = 600$. Replacing the expert distribution sampling used in STS with the Bayesian Aggregation allows us to obtain performances challenging those of the Thompson Sampling Oracle (figure A.2).



(a) (b)

Figure 2: Overall comparison with the state of art and the oracle.

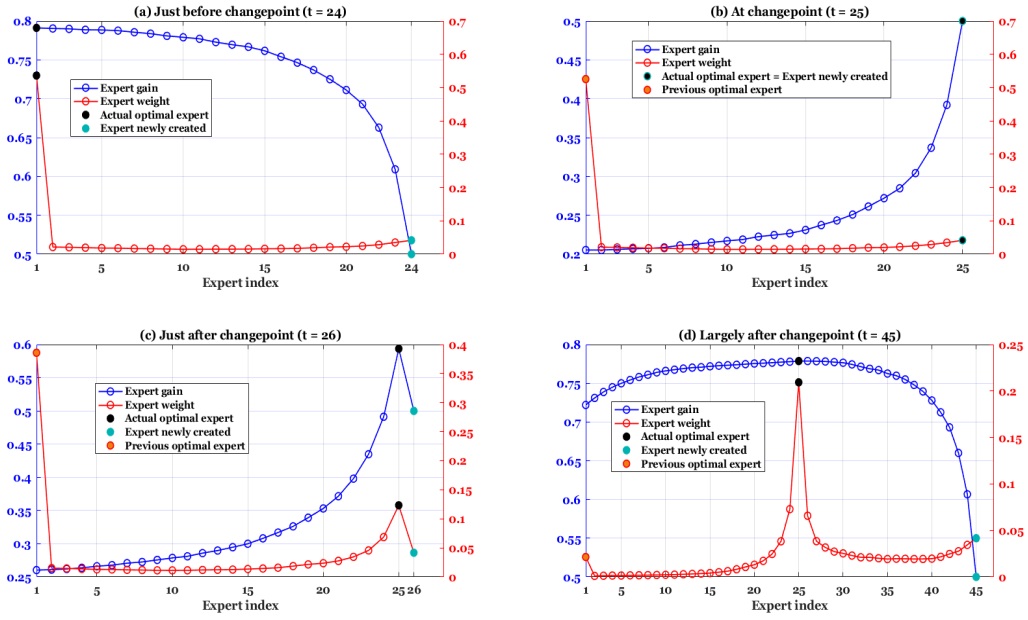
B Numerical Illustrations in the global switching setting

B.1 Tracking the optimal expert in the global switching setting

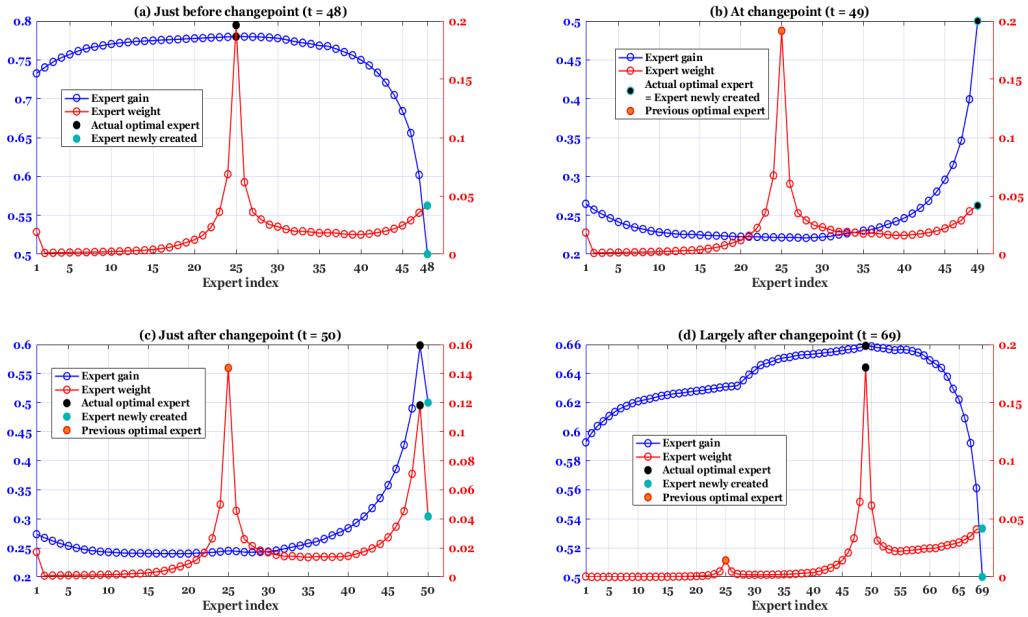
Optimal expert Let $t > 0$ and let $\tau_{[t]}$ denotes the most recent change point before time t. At each time t, a set of exactly t experts is available (each expert has started at $i \in [\![1, t]\!]$). The optimal expert is the one which has started exactly at $t = \tau_{[t]}$. It is the expert which gives the best description of the environment.

Tracking the optimal expert Before the decision step, Global-STS-BA characterizes each arm k by aggregating all the contributions of the available experts. To do well, Global-STS-BA needs to highlight the contribution of the optimal expert because of its optimal description of the environment. This task is called *the track of the optimal expert*.

The estimation of the expert distribution (expert weight) is built according to the instantaneous gain of each expert. This gain takes into account the history of past rewards D_{t-1}. When a switch occurs, the instantaneous gains of all experts fall down because of their obsolete estimation giving the advantage to the expert newly created which will become the optimal expert during the next rounds.
Let us consider a GS-MAB of three arms observing two change points at time $t = 25$ and $t = 49$. Figure B.1 shows the behavior of the expert gain and the expert weight when a switch occurs. It should be noted that the experts are indexed by their starting time i.e. the most recent expert is the one with highest index and inversely.



(a) Tracking the optimal expert during the first switch



(b) Tracking the optimal expert during the second switch

Discussions

- Just before the change point, the optimal expert (index = 1 for the first switch and index = 25 for the second one) is given the highest weight. The lower the index, the higher the expert gain. This behavior is expected as much as the higher the number of the observations the higher the expert gain.

- When the switch occurs, all the experts created before the change point see their gains dropping sharply because of the abrupt change of the environment i.e. the instantaneous gain doesn't match anymore with the current environment, except the expert newly created which is given a gain based on the prior of a Thompson Sampling $\left(\frac{\alpha_0}{\alpha_0+\beta_0}\right)$.
- Just after the change point, the new optimal expert see its weight starting to grow up while the weight of the previous optimal expert starts to fall down.
- Largely after the change point, the optimal expert is given the highest weight, because of its well fit of the current environment. This is corresponding to the expected behavior of the Global-STS-BA: **the optimal expert is well tracked.**

B.2 Behavior of Global-STS-BA with respect to the memory size

Regarding the implementation of the Global-STS-BA, we should notice that memory space and time requirements of the experts inference model grow linearly at each time step because of the creation of a new expert. This makes the size of the support set of the expert distribution $\{w\}_t$ increasing by one. Thus, for computational reasons, we propose to restrict the number of experts by fixing a maximum memory size M. So, at each round, after launching the inference model, we delete the worst expert $f_t^{min} = \arg\min_i w_{i,t}$. In all the experiments, we consider a GS-MAB of three arms. Changes occur at each 1000 rounds. We variate the memory size from $M = 15$ to $M = 3000$. Experiments are run 100 times.

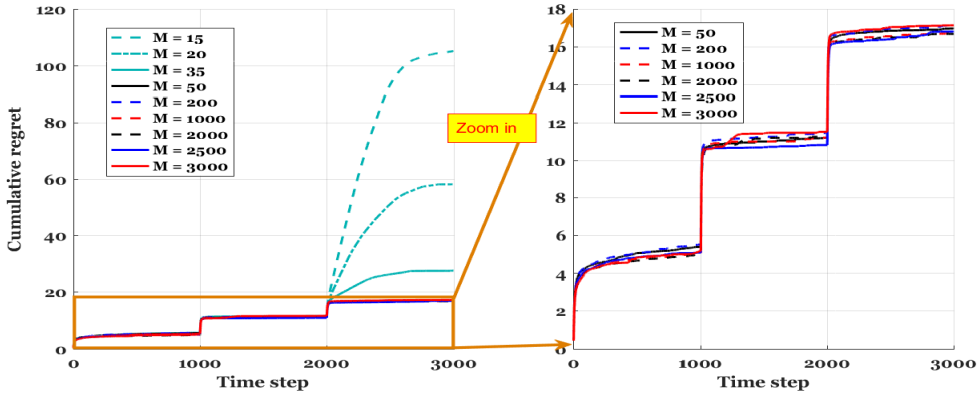
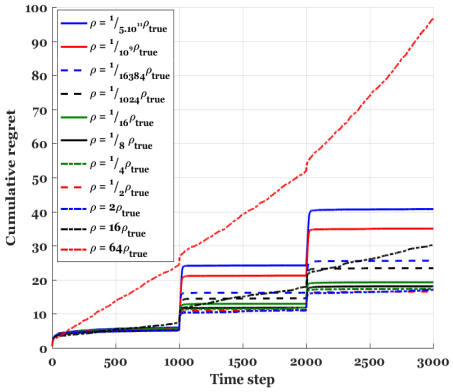


Figure 3: Behavior with respect to the switching rate ρ. For a not very poor memory size value ($M \geqslant 50$), the performances of the Global-STS-BA remains stable.

B.3 Behavior of Global-STS-BA with respect to the switching rate ρ

We should also notice that the inference model of the experts needs the knowledge of the true switching rate (ρ_{true}). In real life, this value is unknow to the agent. [16] has proposed to learn the switching rate from the data via a gradient descent. This method appears to not perform particularly well if the switching rate has to be adapted at every time step. Figure B.3 shows the behavior of the Global-STS-BA for different values of the switching rate.



Discussions The performances of the Global-STS-BA remains stable even if the switching rate is quite far from the true one.

Algorithm 4 Per-Arm Switching **T**hompson **S**ampling with **B**ayesian **A**ggregation

procedure PER-ARM-STS-BA($\mathcal{K}, T, \alpha_0, \beta_0, \rho$)
  $t \leftarrow 1$, and $\forall k \in \mathcal{K}\ w_{k,1,t} \leftarrow 1, \alpha_{k,1,t} \leftarrow \alpha_0,\ \beta_{k,1,t} \leftarrow \beta_0$ ▷ Initializations
  for $t \leqslant T$ **do** ▷ Interaction with environment
    $k_t \leftarrow$ CHOOSEARM $(\{w\}^t, \{\alpha\}^t, \{\beta\}^t)$
    $x_{k_t} \leftarrow$ PLAYARM(k_t) ▷ Bernoulli trial
    $\{w\}_{t+1} \leftarrow$ UPDATEEXPERTWEIGHT $(\{w\}_t, \{\alpha\}_{k_t,t}, \{\beta\}_{k_t,t}, x_{k_t}, \rho)$
    $\{\alpha\}_{t+1}, \{\beta\}_{t+1} \leftarrow$ UPDATEARMMODEL$(\{\alpha\}_t, \{\beta\}_t, x_{k_t}, k_t)$

procedure CHOOSEARM$(\{w\}_t, \{\alpha\}_t, \{\beta\}_t)$
  $\forall\ k \in \mathcal{K}\ \forall\ i \in [\![1,t]\!]\ \ \theta_{k,i,t} \leftarrow \mathrm{Beta}(\alpha_{k,i,t}, \beta_{k,i,t})$
  return $\arg\max_k \sum_{i \in [\![1,t]\!]} \frac{w_{k,i,t}}{\sum_{j \in [\![1,t]\!]} w_{k,j,t}}\, \theta_{k,i,t}$ ▷ Bayesian aggregation

procedure UPDATEEXPERTWEIGHT$(\{w\}_t, \{\alpha\}_{k_t,t}, \{\beta\}_{k_t,t}, x_{k_t}, \rho)$
  $l_{i,t} \leftarrow -x_{k_t} \log\left(\frac{\alpha_{k_t,i,t}}{\alpha_{k_t,i,t}+\beta_{k_t,i,t}}\right) - (1-x_{k_t}) \log\left(\frac{\beta_{k_t,i,t}}{\alpha_{k_t,i,t}+\beta_{k_t,i,t}}\right) \forall\ i \in [\![1,t]\!]$
  $w_{k_t,i,t+1} \leftarrow (1-\rho)\, w_{k_t,i,t} \exp(-l_{i,t}) \forall\ i \in [\![1,t]\!]$ ▷ Increasing the size of expert $f_{k_t,i,t}$
  $w_{k_t,t+1,t+1} \leftarrow \rho \sum_i w_{k_t,i,t} \exp(-l_{i,t})$ ▷ Creating new expert for arm k_t starting at $t+1$
  $w_{k,i,t+1} \leftarrow (1-\rho)\, w_{k,i,t} \forall\ i \in [\![1,t]\!]$ and $\forall\ k \neq k_t$ ▷ Increasing the size of expert $f_{k,i,t}$
  $w_{t+1,t+1,k} \leftarrow \rho \sum_i w_{k,i,t}\ \forall\ k \neq k_t$ ▷ Creating new expert for arm k starting at $t+1$
  return $\{w\}_{t+1}$

procedure UPDATEARMMODEL$(\{\alpha\}_t, \{\beta\}_t, x_{k_t}, k_t)$
  $\alpha_{k_t,i,t+1} \leftarrow \alpha_{k_t,i,t} + \mathbb{1}(x_{k_t} = 1)\ \forall\ i \in [\![1,t]\!]$
  $\beta_{k_t,i,t+1} \leftarrow \beta_{k_t,i,t} + \mathbb{1}(x_{k_t} = 0)\ \forall\ i \in [\![1,t]\!]$
  $\alpha_{k,t+1,t+1} \leftarrow \alpha_0,\ \ \beta_{k,t+1,t+1} \leftarrow \beta_0\ \ \forall\ k \in \mathcal{K}$ ▷ Initializing new expert
  return $\{\alpha\}_{t+1}, \{\beta\}_{t+1}$